Research Journal of Information Technology 2(2): 39-42, 2010
ISSN: 2041-3114

Submitted date: August 10, 2010 Accepted date: September 13, 2010 Published date: November 15, 2010

Computing Rule Confidence using Rough set and Data Mining

P. Ramasubramanian, V. Sureshkumar, S. Nachiyappan and C.B. Selvalakshmi
Department of CSE, Velammal College of Engineering and Technology,
Madurai - 9, Tamilnadu - India

Abstract: Rough Set theory is a new mathematical tool to deal with representation, learning, vagueness, uncertainty and generalization of knowledge. It has been used in machine learning, knowledge discovery, decision support systems and pattern recognition. It can abstract underlying rules from data. Confidence is the criterion to scaling the reliability of rules. Traditionally, the algorithm to obtain the deduction of decision rule in rough sets theory always take more into account of the number of decision rules than the cost of the rules. In this study, we reconstruct the formulae for CF_1 and CF_2. Further, the study is the scope of placement of students based on three input parameters, which considers effect on confidence caused by both imperfect and incompatible information.

Key words: Confidence, data mining, decision support, geometric mean, knowledge discovery, perfect and imperfect information, rough set theory

INTRODUCTION

Pawlak, (1991) a Polish mathematician, put forward Rough Sets Theory (RST) in 1980, which is a data reasoning method. In recent years, it has been rapidly developed in the field of intelligent information processing. This theory deals with representation, learning and generalization of uncertain knowledge (Huanglin, 1996; Jia-Cheng *et al*., 2003; Jianguo, 2002; Pawlak, 1991), and its distillate is to reduce knowledge and produce concise decision rules without any prior information beyond the data set to be dealt with.

One of the main topics of rough set theory is extraction of optimal, consistent decision rules from decision tables. RST, belonging to a kind of non-determinstic reasoning methods, abstracts decision rules from an uncertain data set. Confidence is the criterion to scaling the reliability of rules. It avoids the blindness when we employ these decision rules. So it is significantly important to compute confidence. Ye (2002) put forward a formula to compute confidence when RST deals with imperfect information, but it is rather deficient. The following text will synthesize and improve these methods, and give a novel and relatively complete approach to computing confidence for RST rules. This approach is based on the principles of RST, computes confidence from data.

COMPUTING METHOD FOR RULE CONFIDENCE

Basic idea: The information system in RST is generally expressed in terms of a decision table. Data in the table are got through observation and measurement. Its rows represent objects, and its columns represent attributes divided into two sets, i.e., a condition attributes set and a decision attributes set. RST reduces the decision table, and then produces minimum decision rules sets. Now we adopt C-F model which is a basic method of confidence theory in the field of artificial intelligence. The general form of RST rules can be expressed by:

RULE : IF C THEN D (CF(D,C))

where 'C' represents conditional attributes and their values, 'D' represents a decision attributes and their values, 'CF(D,C)' represents confidence of the rule, also called credit factor, which describes the reliability of the statement that D is true when C is true. CF (D, C) belongs to [0, 1]. The bigger it is, the more reliable the rule is.

Suppose that the measurement is exact, the uncertainty of information in a decision table mainly includes two parts: imperfection and incompatibility. Here imperfection represents that some kinds of test data pattern set do not appear in the table but they exist theoretically. Incompatibility represents that some objects have different decision classes' while they have the same precondition. The concept is related to indiscernibility: The imperfection and incompatibility of information in the decision table cause that RST rules are not completely reliable, i.e. the confidence is less than 1.

Suppose that CF_1 denotes the confidence caused by information imperfection, and CF_2 denotes, the confidence caused by information incompatibility, RST rule confidence is defined by:

Corresponding Author: P. Ramasubramanian, Department of CSE, Velammal College of Engineering and Technology, Madurai - 9. Tamilnadu - India

$$CF(D,C) = CF_1(D,C) * CF_2(D,C) \quad (1)$$

If we can calculate CF_1 and CF_2, then we get CF(D,C) The effects on different decision rules, caused by the imperfect and incompatible information in the decision table, are significantly different, so we have to calculate CF_1 and CF_2, respectively for each rule.

Computing CF_1: If a decision table includes n attributes, and an attribute has m_i (i = 1, 2... n) values, then, without regard of incompatible information, the decision table is perfect when there are $\prod m_i$ patterns abstracted from objects. Otherwise the decision table is imperfect. Ye (2002) defined the rule confidence based on information imperfection as follows:

CF_1 (RULE: C → D) =
Card (perfect pattern set)/Card (imperfect pattern set) (2)

where the numerator represents the number of all the patterns in the perfect set which have appeared in the decision table and met the condition C, the denominator represents the number of all the patterns in the imperfect set which should appear in the decision table theoretically and meet the condition C.

However CF_1 calculated from this equation is generally less than real CF_1, sometimes greatly less than the real value, even less than the value when the rule is got without any data. Therefore, Eq. (2) needs to be modified. The modified equation is as follows:

$$CF_1 (RULE: C \rightarrow D) = 1/n (a + 4b + c / 6) \quad (3)$$

where 'a' represents the number of all the patterns in the imperfect set which meet the condition C and have already appeared in the decision table, 'b' represents the number of all the patterns in the perfect set which meet the condition C and should appear in the decision table, 'c' represents the sum decision classes attributes for the particular input parameters and 'n' represents the total number of attributes in Table 1.

Computing CF_2: If in a decision table objects with the same precondition have different decision, information incompatibility is caused. Now consider RULE: if C then D. According to rough membership function, rule confidence based on information incompatibility is defined as follows:

$$CF_1 (RULE: C \rightarrow D) = \text{Card} (Y \cap [x]_R) / \text{Card} ([x]_R) \quad (4)$$

where 'Y' represents the set of the objects which meet decision D in the decision table, '$[X]_R$' represents the set of the objects which meet condition C in the table, 'R' is the attributes set related to condition C, 'Card( )' represents cardinal number of the set.

Table 1: Student placement information

Id	English	Finance	Personality	Placement
1	Yes	Low	Medium	Medium
2	No	Low	Low	Low
3	No	Medium	High	Medium
4	Yes	High	Low	Low
5	No	Medium	Low	Low
6	Yes	Medium	Medium	Medium
7	No	Low	High	Medium
8	Yes	Medium	High	High
9	Yes	Medium	Low	Medium
10	No	High	Medium	High

Computing CF: CF_1 can be calculated by Eq. (3), CF_2 by Eq. (4), and then we can get CF (D, C) by Eq. (1). When the information related to the given rule in the decision table is perfect and compatible, the confidence of the rule is equal to 1. Otherwise it is less than 1.

An example: Let us consider the student placement information decision in Table 1.

In Table 1, there are 10 objects. The set of condition attributes is {English Medium, Financial Status, and Personality Development}, the set of decision attributes is {Placement}. The value of the English Medium has two values namely Yes or No. The value of the Financial Status is calculated on the basis of Parents monthly income and additional income. The value of the Personality Development is calculated on the basis of student performances in Subject depth, Communication skills, Participating seminar, Public, social activities, and Co-curricular activities. Now, we wish to analyze the following sample information table using rule confidence in rough set theory.

The decision attribute 'Placement' has three values, i.e., three decision classes: Low, Medium, and High. The three condition attributes have 2, 3 and 3 kinds of values respectively, i.e., $m_1 = 2$, $m_2 = 3$, $m_3 = 3$. So the number of a perfect set is $\prod m_i = 2 * 3 * 3 = 18$. *So* we can deduce that Table 1 is imperfect.

In the Table 1, Object 1 and 4 are completely the same. Object 9 has the same condition attributes and values, but they have different decision class, so the object is incompatible.

Now, we will calculate CF_1 and CF_2 by using 3 and 4 as follows:

According to Table 1, the rules produced by RST and their confidence are as follows:

$$R_1: (\text{Personality, High}) \rightarrow (\text{Placement, High})$$

In the Table 1, objects meeting with condition (Personality, High) are the 3rd, 7th and 8th rows, of which the 8th row has the same condition attributes and values.

Then we can conclude that the number of the patterns in the perfect set is 1; which meets condition (Personality, High), and the number of patterns in the imperfect set is 2. The number of kinds of decision classes for input parameter is 3. Therefore:

$$CF1\ (R_1) = 1/3\ ((2+4*1+3)/6) = 1/3\ (9/6) = ½ = 0.5$$

Now, we need to calculate CF_2. In the table objects with decision (Placement, High) are the 8^{th} and 10^{th} rows. So:

$CF_2\ (R_1)$ = Card ({8, 10} ∩ {3, 7, 8})/Card {3, 7, 8}
= Card {8}/ Card {3, 7, 8} = 1/3 = 0.33

Then the confidence of R_1 equals:

$CF(R_1) = CF_1 * CF_2 = 0.5 * 0.33 = 0.17$
R_2: (Personality, High) → (Placement, Medium)
$CF_1(R_2) = 1/3\ ((2+4*2+3)/6) = 1/3\ (13/6) = 13/18 = 0.72$
$CF_2\ (R_2)$ = Card ({1, 3, 6, 7, 9} ∩ {3, 7, 8})/Card {3, 7, 8}
= Card {3, 7}/ Card {3, 7, 8} = 2/3 = 0.67

Then the confidence of R_2 equals:

$CF(R_2) = CF_1 * CF_2 = 0.72 * 0.67 = 0.48$
R_3: (Personality, Medium) → (Placement, High)
$CF_1(R_3) = 1/3\ ((1+4*1+3)/6) = 1/3\ (8/6) = 4/9 = 0.44$
$CF_2\ (R_3)$ = Card ({8, 10} ∩ {1, 6, 10})/Card {1, 6, 10}
= Card {10}/ Card {1, 6, 10} = 1/3 = 0.33

Then the confidence of R_3 equals:

$CF(R_3) = CF_1 * CF_2 = 0.44 * 0.33 = 0.15$
R_4: (Personality, Low) → (Placement, Medium)
$CF_1(R_4) = 1/3\ ((4+4*1+3)/6) = 1/3\ (11/6) = 11/18 = 0.61$
$CF_2\ (R_4)$ = Card ({1, 3, 6, 7, 9} ∩
{2, 4, 5, 9})/Card {2, 4, 5, 9}
= Card {9}/Card {2, 4, 5, 9} = 1/4 = 0.25

Then the confidence of R_3 equals:

$CF(R_4) = CF_1 * CF_2 = 0.61 * 0.25 = 0.15$
R_5: (Finance, Low) and (Personality, Low) →
(Placement, Low)

In the Table 1, objects' meeting with the condition of R5 is 2^{nd} only. The set-of related patterns appearing in this table is perfect.

So $CF_1(R_5)$ equals 1.

The objects meeting R_2 are all compatible, so CF_2 (R_5) also equals 1.

Then we can get the confidence of rule R_5, CF (R_5) = 1, i.e., R_5 is thoroughly reliable.

R_6: (Finance, Low) and (Personality, Medium) →
(Placement, Medium)

$CF_1(R_6) = 1/3\ ((4+4*1+6)/6) = 1/3 \times 14/6 = 7/9 = 0.77$
$CF_2\ (R_6)$ = Card ({1, 3, 6, 7, 9} ∩ {1})/Card {1}
= Card {1}/ Card {1} = 1/1 = 1
$(R_6) = CF_1(R_6) * CF_2(R_6) = 0.77 * 1 = 0.77$

R_7: (Finance, Medium) and (Personality, High) →
(Placement, High)
$CF_1(R_7) = 1/3\ ((1+4*1+6)/6) = 1/3 \times 11/6 = 11/18 = 0.61$
$CF_2\ (R_7)$ = Card ({8, 10} ∩ {3, 8})/Card {3, 8}
= Card {8}/ Card {3, 8} = 1/2 = 0.5
$CF\ (R_7) = CF_1\ (R_7) * CF_2\ (R_7) = 0.61 * 0.5 = 0.31$

R_8: (Finance, High) and (Personality, Medium) →
(Placement, High)
$CF_1(R_8) = 1/3\ ((1+4*1+6)/6) = 1/9 \times 11/6 = 11/54 = 0.20$
$CF_2\ (R_8)$ = Card ({8, 10} ∩ {10})/Card {10}
= Card {10}/ Card {10} = 1/1 = 1
$CF(R_8) = CF_1(R_8) * CF_2(R_8) = 0.20 * 1 = 0.20$

R_9: (Finance, High) and (Personality, Low) →
(Placement, Low)

In the Table 1, objects' meeting with the condition of R9 is 4^{th} only. The set-of related patterns appearing in this table is perfect.

So $CF_1(R_9)$ equals 1.

The objects meeting R_2 are all compatible, so CF_2 (R_9) also equals 1.

Then we can get the confidence of rule R_5, CF (R_9) = 1, i.e., R_9 is thoroughly reliable.

R_{10}: (English, Yes) and (Finance, Medium)
and (Personality, Medium) → (Placement, Low)
$CF_1(R_{10}) = 1/3\ ((4+4*1+8)/6) = 1/3 \times 16/6 = 16/18 = 0.89$
$CF_2\ (R_{10})$ = Card ({1, 3, 6, 7, 9} ∩ {6})/ Card {6}
= Card {6}/ Card {6} = 1/1 = 1
$CF(R_{10}) = CF_1(R_{10}) * CF_2(R_{10}) = 0.89 * 1 = 0.89$

In the above discussion, the effect caused by information imperfection and computing result is more appropriate.

CONCLUSION

In summary, rough sets theory is a new method for analyzing, inducing, studying and discovering of data. It puts forward a method of student placement assistance based on rough set theory. According to the above discussion, we conclude that the approach put forward in this study to compute rule confidence in rough sets theory entirely considers the imperfection and incompatibility of

the information in the decision table. It is a relatively appropriate method. It can calculate rule confidence based on given data and avoid subjectivity and one-sidedness. Its signification is clear, and it is easy to compute and apply. Based on this approach, we can further consider the case of the preconditions with credit factors, and compute the confidence of rule combination.

ACKNOWLEDGMENT

We acknowledge Prof. N. Sathish Nainar, Professor, Department of mathematics, Anna University, Tirunelveli for his help in collecting the data required for the analysis.

REFERENCES

Huanglin, Z., 1996. Rough Sets Theory and its Applications. Chongqing University Press, Cbongqing.

Jia-Cheng, W., D. Hua-Ping and S. You-Xian, 2003. A novel approach to computing rule confidence in rough set theory. Proceedings of the 2nd International Conference on Machine Learning and Cybernetics, Xi'an, 2-5 Nov., pp: 1523-1526.

Jianguo, T., 2002. Reliability analysis to deal with imperfect information by rough set theory. Control Decis., 17(2): 255-256.

Pawlak, Z., 1991. Rough Set-Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

Ye, D., 2002. Computation of credit factors for rules and inference based on rough membership function. J. Fuzhou University (Natural Science Edn.), 30(3): 294-297.